





Study on the detection method of biological characteristics of hepatoma cells based on terahertz time-domain spectroscopy

HANXIAO GUAN,¹ WEIHANG QIU,² HENG LIU,¹ YUQI CAO,^{1,*} 
LIANGFEI TIAN,² PINGJIE HUANG,¹ DIBO HOU,¹  AND GUANGXIN ZHANG¹

¹State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, 310000, China

²College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, 310000, China

*yuqicao@zju.edu.cn

Abstract: Liver cancer usually has a high degree of malignancy and its early symptoms are hidden, therefore, it is of significant research value to develop early-stage detection methods of liver cancer for pathological screening. In this paper, a biometric detection method for living human hepatocytes based on terahertz time-domain spectroscopy was proposed. The difference in terahertz response between normal and cancer cells was analyzed, including five characteristic parameters in the response, namely refractive index, absorption coefficient, dielectric constant, dielectric loss and dielectric loss tangent. Based on class separability and variable correlation, absorption coefficient and dielectric loss were selected to better characterize cellular properties. Maximum information coefficient and principal component analysis were employed for feature extraction, and a cell classification model of support vector machine was constructed. The results showed that the algorithm based on parameter feature fusion can achieve an accuracy of 91.6% for human hepatoma cell lines and one normal cell line. This work provides a promising solution for the qualitative evaluation of living cells in liquid environment.

© 2023 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

In recent years, malignant tumor lesions have become one of the major diseases endangering human health and life [1,2]. Pathological studies show that nearly 90% of deaths in cancer patients are attributed to metastasis [3,4]. Detecting circulating tumor cells (CTC) is of great significance for the real-time monitoring of tumor development, especially for the early detection of cancer. Besides, it can also predict potential recurrence and assess the risk of death [5,6]. As a cutting-edge tumor detection technology, liquid biopsy utilizes human body fluids as a sample source for detection, diagnosis, and in vitro experiments. This technology meets the need for efficient and non-invasive tumor diagnosis and is applicable to both primary and metastatic patients. Therefore, it has attracted extensive attention from scholars worldwide. Compared with traditional methods, liquid biopsy is advantageous for simpler, non-invasive sampling, good repeatability, available for continuous monitoring, and overcoming tumor heterogeneity [7].

Though able to detect circulating tumor cells [8], the practical operation of liquid biopsy is still limited due to the rarity and structural complexity of CTCs. At present, immunostaining is the mainstream method for CTC detection [9–12]. The methods require multi-step cell preparation and extraction process, which may lead to the loss and damage of tumor cells and adversely affect the accuracy of detection. Also, the complete physiological state of living cells, which has a key influence on the subsequent molecular typing, tumor staging and drug action detection

experiments, may be interfered by external labeling chemicals in immunostaining. Therefore, there is an urgent need for a label-free technique that can maintain the cellular properties.

Terahertz (THz) wave is a kind of electromagnetic radiation with a frequency band of 0.1-10 THz (1 THz = 10^{12} Hz), located between microwave and infrared radiation [13]. The vibrational and rotational frequencies of molecular chemical bonds in tumor markers such as nucleic acids, proteins, carbohydrates, and abnormal metabolites are located in the THz band, thus they can be characterized in THz spectrum. Based on the properties of THz waves, THz technology can distinguish the cancer cells from the target cells according to their differences in composition and structure [14–16]. Therefore, THz technology has extreme sensitivity to the morphology and physiological characteristics of cells, which is conducive to evaluating the complete physiological activities of living cells in different periods from a dynamic perspective and realizing long-term monitoring. It is considered to introduce terahertz time-domain spectroscopy (THz-TDS) to study tumor cell lines in liquid environment [17].

Researchers have explored the differential characterization of the hydration state, permeability, and other characteristics of tumor cells, as well as the changes in cytopathologic status induced by drugs or culture medium environment. Shiraga's team used the THz attenuated total reflection system to detect three kinds of tumor cells, DRD-1, HEK293 and HeLa, using the complex dielectric constant derived from the THz spectrum, proving that THz spectroscopy can characterize the dynamic characteristics of water molecules in human tumor living cells [18]. Grognot et al. used saponins to infiltrate epithelial cells and used THz attenuated total reflection imaging to conduct real-time measurement of cytoplasmic leakage. The result is proven to be consistent with those of standard bicinchoninic acid protein assay [19]. Zou et al. detected the dielectric response of mammary epithelial cells and recorded the state changes of cells under oxidative stress. They verified the results of THz detection using fluorescent-labeled optical imaging and flow cytometry, demonstrating that this technique can monitor the process of apoptosis in real time [20]. Various machine learning methods were also used to analyze the detection results [21,22]. Liu et al. screen hepatocellular carcinoma by THz pulse signals combining variational mode decomposition and composite weighted-scale sample entropy. This method can distinguish similar signals [23]. Cherkasova et al. adopted Random Forest (RF) and Extreme Gradient Boosting to analyze the blood plasma samples of glioma patients [24]. Yang et al. combine principal component analysis (PCA) and support vector machine (SVM) to identify benign and malignant cell component [25].

In this study, we cultured two human hepatoma cell lines Huh-7 and HepG2, and a normal cell line MIHA, and then detected the cell lines at four different concentrations while maintaining cell activity. On this basis, we explored the differential characterization of the interaction response between THz waves and different living cells. Absorption coefficient and dielectric loss were selected based on statistical analysis method and correlation analysis. Then, combined with the absorption coefficient and dielectric loss spectra, we employed the maximum information coefficient (MIC) and the principal component analysis to extract the frequency domain features, and we adopted the SVM model for identification, so as to realize the qualitative evaluation of living cells.

2. Methods

2.1. Experimental preparation

In this study, a Z-3 THz-TDS system developed by Zomega Corporation, USA, was employed, and a Vitesse-800-5 mode-locked Ti:sapphire femtosecond laser by Coherent Company of USA was applied as the excitation source. The laser pulse is centered at 800 nm with a pulse width of less than 100 fs and the output power is 960 mW. More details of the system can be found in our earlier work [26].

The liquid sample cell was purchased from Hellma Analytic (Germany), as shown in Fig. 1. The windows of cuvette were made of suprasil. Liquid samples can be injected from the two small holes in the upper part with a pipette. The amplitude ratio and phase difference of the first transmission are very close to the measured value of multiple refractions through experiments, and the error is less than 3%, which verifies that the multiple refractions of the window can be ignored. It has an optical path of 0.1 mm and an inner diameter of 13 mm. The total volume including the pipes is 160 μL . During the measurement, the environment was kept at room temperature, and nitrogen flow was continuously injected into the system to keep the humidity below 1%. Due to the downward trend of transmission coefficient, 0.2-1.4 THz was applied as the effective frequency range in this paper.

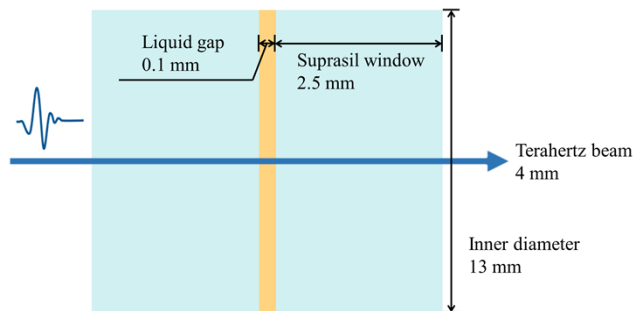


Fig. 1. Schematic diagram of intersecting surface of the liquid cell.

Hepatoma cell lines Huh-7, HepG2 and normal liver cell lines MIHA were cultured using DMEM medium (supplemented with 10% fetal bovine serum, 50 $\mu\text{g/mL}$ streptomycin and 50 $\mu\text{g/mL}$ penicillin) in an incubator with 5% carbon dioxide concentration at 37 $^{\circ}\text{C}$. The pre-cultured cells were digested by 0.25% trypsin/EDTA, which was removed by centrifugation at 1200 rpm for 5 min, and the concentrated cells were re-dispersed in the medium (1 mL). Considering the needs for cell growth and for the characterization of cellular properties of various concentrations, we took twelve different cell suspensions with four concentrations of 10^3 cell/mL, 10^4 cell/mL, 10^5 cell/mL, and 10^6 cell/mL.

Before the experiment, the cells were evenly distributed in the culture by blowing with a pipette gun and were then injected into the liquid pool. After the detection was completed, the residual cells and proteins were cleaned with aqua regia, then washed with distilled water, and the residual water was blown out by an air pump to make the inside fully dry. During the detection, we ensured that the cell activity was always maintained above 85%. In addition, by rotating the liquid pool and changing its angle with the direction of the electric field intensity of the THz wave, it was proved that the parameters, such as peak value, time delay, refractive index, and absorption coefficient, were basically independent of the sample's orientation, meaning the cell suspension could be regarded as isotropic. 18 samples were taken from each cell suspension, and each sample was measured five times, so a total of 1080 sets of time-domain spectral data were obtained from cell suspensions.

Fast Fourier transform (FFT) was used to calculate the THz frequency domain spectrum from the obtained time domain spectrum. Since the transmission system was a window-liquid-window multilayer structure, the material parameter extraction method proposed by Duvillaret et al. [27] is employed. Considering that the strong absorption of THz wave by liquid can annihilate the micro signals, it has been proved experimentally that the extremely weak signals caused by multiple refractions can be ignored to simplify the THz optical parameter model. Taking the signal of the empty liquid cell as the reference signal and the signal with cell suspension as the sample signal, the refractive index, absorption coefficient, dielectric constant, dielectric loss, and

dielectric loss tangent of the sample are calculated, as shown in Eq. (1) to (5):

$$n(\omega) = \frac{\varphi(\omega)c}{\omega d} + 1 \quad (1)$$

$$\alpha(\omega) = \frac{2k(\omega)\omega}{c} = \frac{2}{d} \ln \left\{ \frac{n(\omega)[1 + n_1(\omega)]^2}{\rho(\omega)[n(\omega) + n_1(\omega)]^2} \right\} \quad (2)$$

$$\operatorname{Re} \left\{ \tilde{\varepsilon}(\omega) \right\} = n(\omega)^2 - k(\omega)^2 \quad (3)$$

$$\operatorname{Im} \left\{ \tilde{\varepsilon}(\omega) \right\} = 2n(\omega)k(\omega) \quad (4)$$

$$\tan \delta = \frac{\operatorname{Im} \left\{ \tilde{\varepsilon}(\omega) \right\}}{\operatorname{Re} \left\{ \tilde{\varepsilon}(\omega) \right\}} \quad (5)$$

where d , c , ω stand for the optical path of the liquid cell, speed of light, and frequency. $n_1(\omega)$ is the refractive index of the window material quartz, $\tilde{\varepsilon}(\omega)$ is complex dielectric constant, whose real part is dielectric constant and imaginary part is dielectric loss. $\tan \delta$ are dielectric loss tangent, $\varphi(\omega)$ and $\rho(\omega)$ are the phase difference and the amplitude ratio between the reference signal and the sample signal, respectively.

2.2. Class separative criterion

To find a set of features that are most effective for classification, it is necessary to measure the effectiveness of classification performance by category separability criteria. At present, the classification separability criterion of geometric distance and probability density is an important basis for classification and discrimination. Ideally, the samples of the same class generally show a densely clustering state in the feature space because of their commonness, while different classes are scattered. Therefore, the inner-class distance between the samples should be smaller than their inter-class distance. In the case of overlapping samples, the feature with a large inter-class distance and a small intra-class distance in the feature space should be selected. Generally, the within-class scatter matrix S_W and the between-class scatter matrix S_B are employed to measure the distance:

$$S_W = \sum_{i=1}^M P(\Omega_i) \frac{1}{N_i} \sum_{k=1}^{N_i} (X_k^{(i)} - m^{(i)})(X_k^{(i)} - m^{(i)})^T \quad (6)$$

$$S_B = \sum_{i=1}^M P(\Omega_i)(m^{(i)} - m)(m^{(i)} - m)^T \quad (7)$$

where M stands for the number of classes, Ω_i stands for the sample set $\{X_1^{(i)}, X_2^{(i)}, \dots, X_{N_i}^{(i)}\}$, $P(\Omega_i)$ stands for the proportion of the sample set in the total set, N_i is the number of samples in one class, m is the mean value of population, and $m^{(i)}$ is the mean value of one class. $\operatorname{tr}(S_W)$ represents the average measure of the characteristic variance of all classes, $\operatorname{tr}(S_B)$ represents a measure of the average distance between the mean value of each class and the mean value of population. We define $J = \operatorname{tr}(S_B)/\operatorname{tr}(S_W)$ as the criterion to represent the separability of classes.

2.3. Maximal Information Coefficient (MIC)

Being able to effectively capture linear, nonlinear, and non-functional association among variables in high-dimensional data sets [28], MIC, as a measure of evaluating feature goodness and redundancy, is widely used in genomics, medical data analysis and other fields [29,30]. Developed on the basis of mutual information in information theory, MIC corrects mutual

information value through unequal interval optimization, making the selected information more detailed, accurate, general and equitable.

The basic principle of MIC is dividing the data set into different grids, then calculating the mutual information among variables according to the distribution of data in the grid. After normalization, the maximum value of mutual information is selected as the final value to approximate the correlation between variables. Assume that variable X represents a feature of cells in THz spectrum and Y represents different cell lines. For the two-dimensional scatter diagram composed of variables, specific grid division is carried out according to different division numbers and positions, then mutual information values are calculated, respectively. For example, divide m and n regions on the x-axis and y-axis, respectively, to make a $m \times n$ grid. Define the grid as G , then the mutual information $I(D|_G)$ of data set $D(X, Y)$ under this division is shown in Eq. (8):

$$I(D|_G) = \sum_{m \in X, n \in Y} p(m, n) \log \left(\frac{p(m, n)}{p(m)p(n)} \right) \quad (8)$$

where $p(m, n)$ is the joint probability distribution of variables X and Y , $p(m)$ and $p(n)$ are the marginal distribution, which is estimated by the probability of data falling in the grid. Since the position of grid division may be equal or unequal, the maximum value of each division is taken as the mutual information value under this division. It is standardized for subsequent comparison, expressed as $M(D)_{m,n}$.

$$M(D)_{m,n} = \frac{\max I(D|_G)}{\log(\min\{m, n\})} \quad (9)$$

Then traverse the values of different m and n , find the corresponding $M(D)_{m,n}$ of each group, and obtain the maximum value of all combinations as MIC

$$MIC(D) = \max_{mn \leq B(N)} \{M(D)_{m,n}\} \quad (10)$$

where, $B(N)$ is the upper limit of grid division, which is generally set to 0.6 power of the total number of samples. In this paper, the MIC of each parameter in the 0.2-1.4 THz frequency band is calculated to select features with higher correlation.

2.4. SVM

SVM, a machine learning method based on structural risk minimization principle with good generalization performance, has wide application in the analysis and recognition of THz spectrum [31,32] for its advantages in solving small-sample-scale, nonlinear dataset. In the case of linear indivisibility, nonlinear mapping is used to map samples to high-dimensional space, and then the maximum margin hyperplane is found to make the data linearly separable, as shown in Eq. (11)

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(w^T \cdot \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \quad (11)$$

where w is the normal vector to classification hyperplane in feature space, b is the intercept of hyperplane, $\phi(x_i)$ is mapping function, ξ_i is slack variable and C is penalty factor. Lagrange multiplier method is used to convert the constraint of maximum margin into a dual problem, so

as to optimize the solution [33]

$$\begin{aligned} \max_{\alpha} & \left\{ \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right\} \\ \text{s.t.} & \sum_{i=1}^N \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \end{aligned} \quad (12)$$

where α_i is the Lagrange multiplier, $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ represents the kernel function. In this paper, radial basis function (RBF) is used as the kernel function, as shown in Eq. (13)

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \quad (13)$$

where γ is the kernel parameter. SVM with RBF kernel function usually improves the model performance by adjusting the penalty factor C and the kernel parameter γ , C measures the relationship between the complexity of the support vector and the misclassification rate, and γ measures the contribution of a single sample to the classification. In this paper, the optimal hyperplane is selected by grid search.

3. Results

3.1. Terahertz spectrum of three cell lines

THz responses of three different cells were measured at four concentrations of 10^3 cell/mL, 10^4 cell/mL, 10^5 cell/mL, and 10^6 cell/mL while maintaining cell viability. The data of 10^5 cell/mL was shown in Fig. 2, where the empty cuvette signal (filled with air) was taken as a reference. It can be seen that there are certain differences in the time delay and the peak value among hepatoma cell lines Huh-7, HepG2, and normal cell line MIHA. To further analyze the characteristics of three cell lines at the concentration of 10^5 cell/mL, refractive index, absorption coefficient, dielectric constant, dielectric loss, and dielectric loss tangent were calculated. The results are shown in Fig. 3.

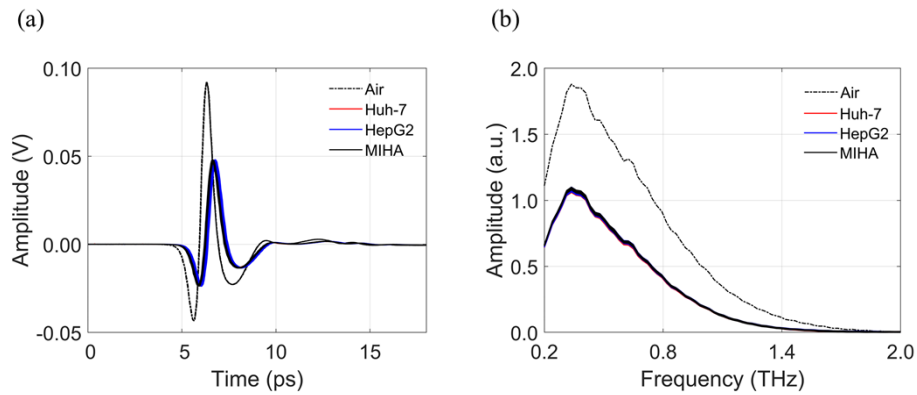


Fig. 2. Terahertz signals and frequency domain spectra for different kinds of cells at the concentration of 10^5 cell/mL. (a) THz time domain spectra. (b) THz frequency domain spectra.

Absorption coefficient and dielectric loss show different responses at some frequencies, as shown in Fig. 3(b) and Fig. 3(d) (the insets show a clearer zoom-in result), respectively. The error bars correspond to the standard deviations. We observe that the results for three kinds of cells do not overlap in some frequency bands of 0.20-0.25 THz. These differences can be employed to

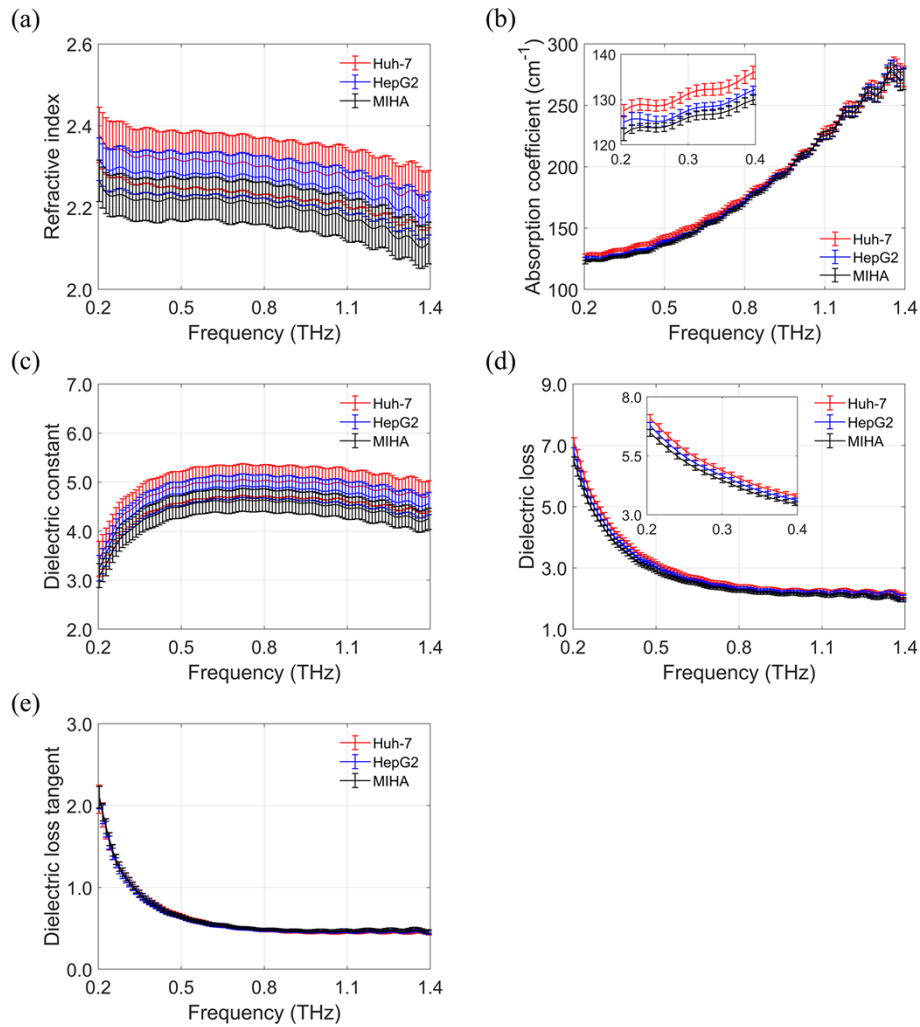


Fig. 3. Spectra of THz frequency parameters for different kinds of cells at the concentration of 10^5 cell/mL. (a) refractive index. (b) absorption coefficient. (c) dielectric constant. (d) dielectric loss. (e) dielectric loss tangent.

better distinguish the three kinds of cells. For refractive index, dielectric constant, and dielectric loss tangent, though their THz responses are different in the frequency band of 0.2-1.4 THz, the error bars are partially overlapped, which make them unideal for the distinguishment. Therefore, the influence of different cells at different concentrations on the interaction between cells and THz waves can be identified by analyzing the spectra of characteristic parameter. Due to the complicated response mechanisms of different characteristic parameters toward the biological features of cells, such as water content and biomarker content, there are differences in the discrimination of each parameter. It is necessary to further analyze such differences and select the parameters that can better identify the tumor cells.

3.2. Feature parameter selection

In order to select parameters that can better characterize cells, class separability criterion was employed to evaluate the classification discrimination. In this paper, we need to compare the classification performance of different THz frequency parameters with different units and ranges. Since there were amplitude variation range differences among the parameters, data scaling was performed through data standardization. Due to the great difference in the trend and curvature of parameters, it was impossible to compare the difference of a certain frequency band directly, as within-class and between-class scatter can only identify the separability of one point. Therefore, $J = tr(S_B)/tr(S_W)$ was calculated as class separability criterion to measure the overall discrimination of each parameter in the effective frequency band of 0.2-1.4 THz. Larger value represents smaller intra-class distance and larger inter-class distance at the same scale. Exemplary data at 10^5 cell/mL were shown in Table 1. It can be seen that the absorption coefficient and the dielectric loss are parameters with high differentiation. The intra-class distances of these two parameters are larger than their inter-class distances, which makes the overlap between different classes as small as possible. This result is consistent with the conclusion obtained from the parameter spectra.

Table 1. Classification separability results of single THz frequency parameter.

Feature parameter	Refractive index	Absorption coefficient	Dielectric constant	Dielectric loss	Dielectric loss tangent
$tr(S_B)$	0.53	0.02	0.24	0.03	0.00
$tr(S_W)$	1.07	0.02	0.59	0.01	0.03
$tr(S_B) / tr(S_W)$	0.50	1.03	0.41	2.14	0.04

To further visualize the results of classification separability, the first two principal components (PC) of parameters were characterized by PCA. The differences in characteristic parameters of cell lines at four concentrations result not only from the influence of cell concentration, but also from the influence of the hydration layer outside the cells [34]. The original 100-dimensional data of four concentrations were used as the input of PCA to find out the parameters with higher specificity. The distributions of the first two PCs of different cells containing data of all concentrations were shown in Fig. 4. Though any single parameter is not sensitive enough for cells at all concentrations, causing partially overlap of characteristic distributions of different cells, it can be seen that the absorption coefficient still has relatively higher differentiation. Specifically, the absorption coefficient is highly sensitive to the change of cell concentration, resulting in a higher dispersion at low concentration and high concentration on PC1. In Fig. 4(b), the mean values of PC2 for MIHA, HepG2 and Huh7 cells are 0.59, 0.48 and 0.37, respectively, which can basically distinguish the three types of cells. (though the mean values of those for three types of cells are almost the same using the other four parameters).

In addition to judging by class separability, we also calculated the correlation and identified the broad functional relationships among variables, so as to select parameters with strong correlation

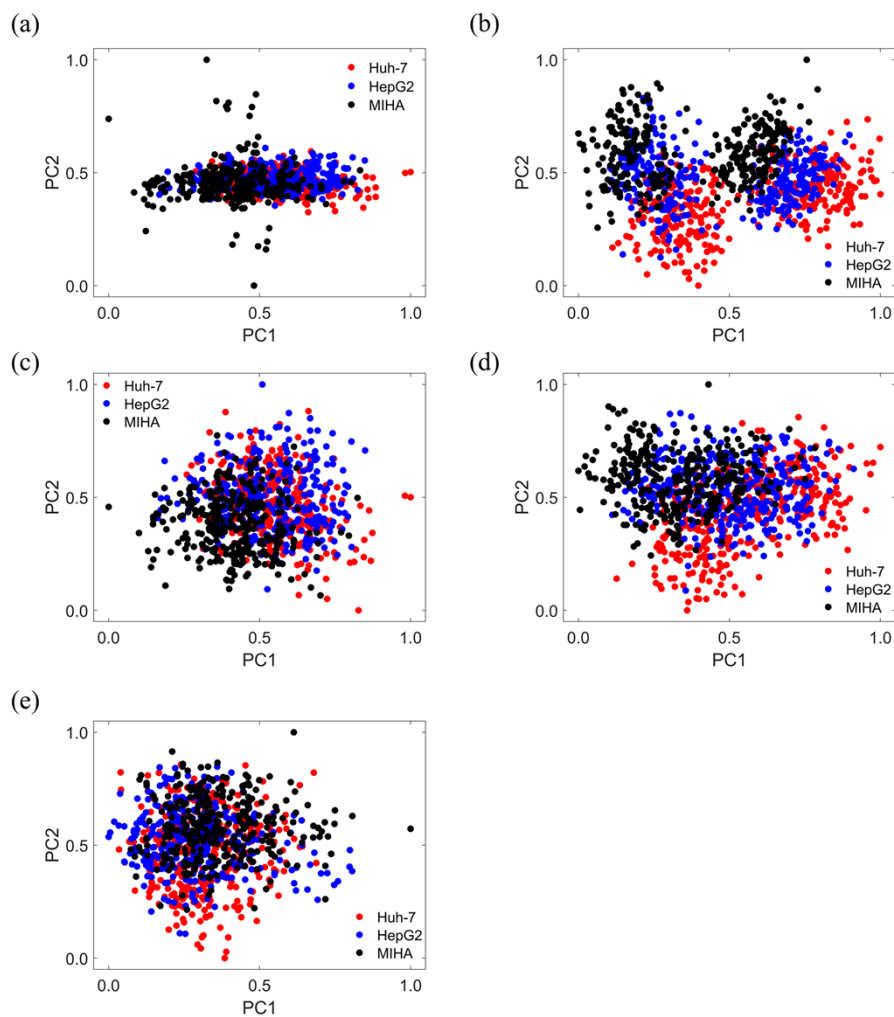


Fig. 4. Visualization of the first two PCs using five parameters. (a) refractive index. (b) absorption coefficient. (c) dielectric constant. (d) dielectric loss. (e) dielectric loss tangent.

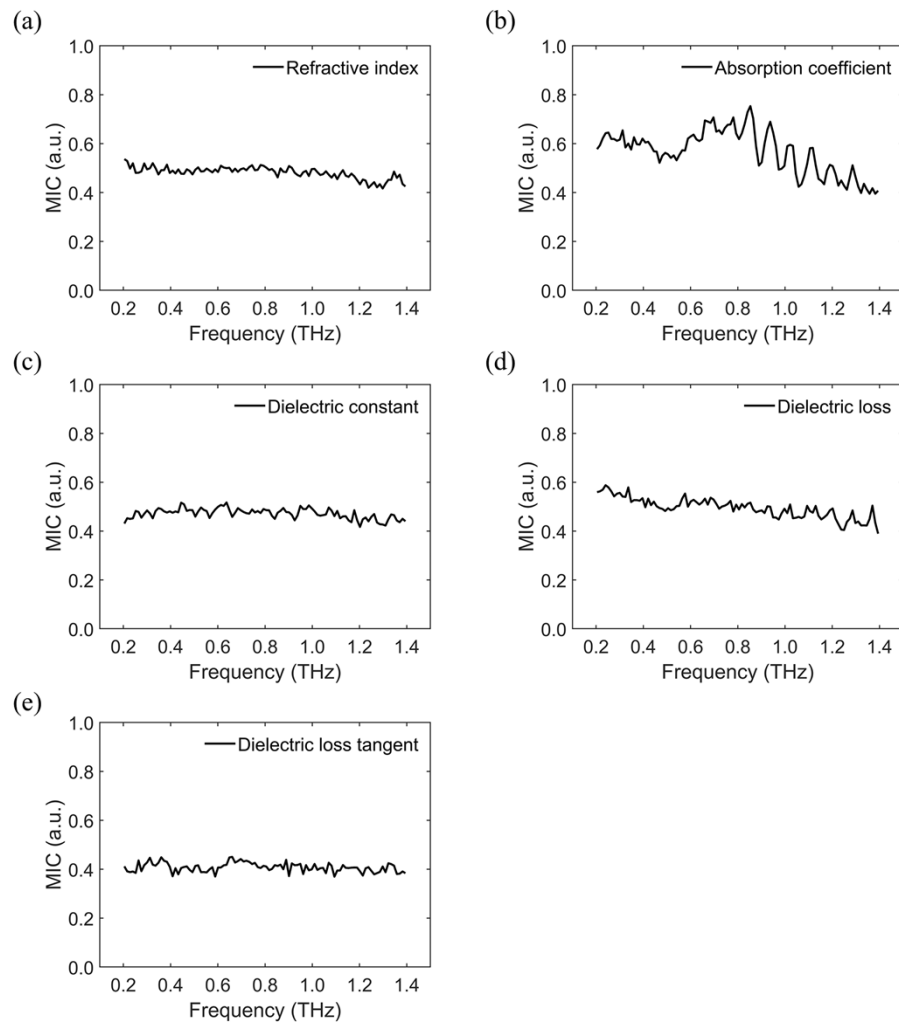


Fig. 5. Maximal information coefficient of THz frequency parameters. (a) refractive index. (b) absorption coefficient. (c) dielectric constant. (d) dielectric loss. (e) dielectric loss tangent.

and interpretability. Therefore, the data of all concentrations were also mixed and grouped according to the type of cells. MIC of each parameter in the frequency band 0.2-1.4 THz were shown in Fig. 5. Specifically, the trend of the MIC of absorption coefficient has a higher fluctuation than the other four parameters. The MIC in the given range is greater than the mean value, which shows a stronger correlation with the cell type and a more comprehensive and sensitive response to cell components. MIC values of other parameters in overall frequency band are relatively balanced and the score is within the range of 0.4-0.5, while the value of dielectric loss in frequency band of 0.2-0.4 THz is about 0.55, indicating higher sensitivity. To sum up, the absorption coefficient can be considered as the main parameter for distinguishing the type of cells, and we use its fusion with the dielectric loss to obtain higher accuracy. In earlier paper [21], it has been found that the absorption coefficient is a parameter that can effectively distinguish between tumor cells and normal cells. It is generally believed that this is because in addition to water, the absorption of THz waves by other components of cells also exhibits significant differences. For example, some scholars have studied the THz response differences of tumor markers such as DNA methylation and carcinoembryonic antigen at different levels.

3.3. Qualitative identification of cells

PCA was employed to reduce 100-dimensional features within the range of 0.2-1.4 THz to 20-dimensional, and SVM was used to classify the extracted features for different types. Since training with mixed concentration was conducive to better recognition of terahertz characteristics, 80% of the mixed data of all concentrations were selected as the training set and 20% as the test set. Then we compared the performance of the models of the above five parameters and the model of absorption coefficient and dielectric loss fusion, and analyzed the impact of different feature selection methods on classification results according to accuracy, precision and recall, as shown in Table 2.

Table 2. Classification results using different feature selection methods.

Feature parameter	Refractive index	Absorption coefficient	Dielectric constant	Dielectric loss	Dielectric loss tangent	Absorption coefficient + Dielectric loss
Accuracy	91.3%	96.2%	94.1%	95.7%	94.9%	98.7%
Precision	91.2%	96.1%	93.5%	95.6%	94.8%	98.8%
Recall	91.3%	96.1%	93.6%	95.6%	94.8%	98.6%

As can be seen from Table 2, when a single parameter is employed for prediction, the results from using the absorption coefficient or the dielectric loss are better than the other three parameters, which is consistent with the results of parameter selection. Compared with the results of using a single parameter, the results using absorption coefficient and dielectric loss fusion are improved by a certain degree, indicating that the sensitive features extracted by the two parameters do not coincide. Thus, feature fusion can avoid information loss to the greatest extent and improve prediction accuracy through complementary advantages.

In order to further describe the differences of cells at different concentrations and judge the impact of the division method on prediction accuracy, we adopted absorption coefficient and dielectric loss fusion as the basis, and we took the mixed data of three high-concentration cells as the training set. A low concentration of 10^3 cell/mL was predicted to classify the type of cells. After reducing the dimension to 40 dimensions by PCA, the classification model was established using RF, back propagation (BP) neural network, and SVM, as shown in Table 3. In addition, the method of extracting 60 dimensions through MIC and then reducing to 40 dimensions by PCA was also compared, as shown in Table 4.

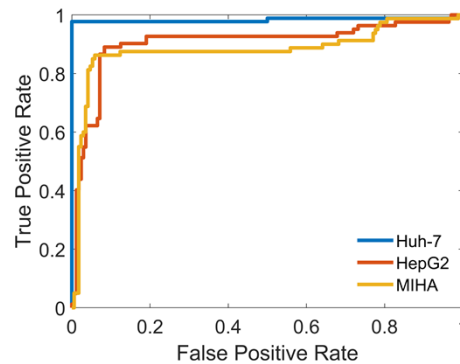
Table 3. Cell qualitative discrimination results using PCA and three classification models.

Algorithm	PCA-RF	PCA-BP	PCA-SVM
Accuracy	76.8%	79.2%	89.2%
Precision	80.0%	87.6%	90.8%
Recall	76.1%	78.8%	89.0%

Table 4. Cell qualitative discrimination results using MIC, PCA and three classification models.

Algorithm	MIC-PCA-RF	MIC-PCA-BP	MIC-PCA-SVM
Accuracy	78.9%	88.3%	91.6%
Precision	82.6%	91.1%	92.3%
Recall	78.9%	88.3%	91.3%

As can be seen from the table, compared with the division method of mixing cells of four concentrations, the prediction accuracy of this method decreases. The characteristics of cells with high concentration are more obvious, while cells with low concentration are more interfered by the water content, as the strong absorption of water to THz waves annihilated part of the effective information. The features extracted from data at higher cell concentrations could not cover the complete characteristics of low concentrations, but the prediction accuracy of SVM still reaches 89.2%. Compared with the other two algorithms, SVM, as a machine learning algorithm suitable for small sample sets, shows a higher prediction accuracy in this data set. Precision and recall are relatively balanced, which reduces the probability of misdiagnosing normal cells as hepatoma cells. The stability and effectiveness of the model are negatively affected by partial noise interference in the THz signal, which may introduce interfering features during PCA feature extraction. Therefore, PCA can be performed better after selecting features with higher correlation through MIC, which can improve the prediction accuracy to 91.6%. Figure 6 shows the ROC curve of the MIC-PCA-SVM model, which has a good recognition ability for all three types of cells, with the best classification performance for Huh-7.

**Fig. 6.** ROC curve of the MIC-PCA-SVM model.

4. Conclusion

In this paper, based on absorption coefficient and dielectric loss fusion, a qualitative identification method of hepatocellular carcinoma living cells based on THz frequency parameters is proposed. Making full use of the THz-TDS, we can calculate the THz parameters which contain abundant

properties. For more accurately extracting pathological characteristics of living cells, we explored the parameters that can better represent cell differences based on the class separability and correlation discrimination. Specifically, based on the class separative criterion of inter-class and inner-class distance, the absorption coefficient and dielectric loss were selected, and visualized by dimension reduction through PCA. Then the MIC was employed to verify the higher correlation between the two parameters. Based on the feature fusion, MIC and PCA were employed for dimension reduction, and SVM was adopted to realize the distinction between two human hepatoma cell lines Huh-7, HepG2 and one normal cell line MIHA. By preferentially selecting strong correlation features and reducing noise interference, accuracy, precision and recall were improved to 91.6%, 92.3% and 91.3%. And it can capture the characteristics of different cells. In the future, we will conduct experiments at lower cell concentrations to improve the detection sensitivity and increase the number of multiple cell combination experiments to further enhance the generalization and stability of the model.

Funding. Key Technology Research and Development Program of Zhejiang Province (2021C03177); State Key Laboratory of Industrial Control Technology Program of Zhejiang University (ICT2023A10); National Natural Science Foundation of China (61873234).

Acknowledgments. We thank Yiwen. E for helpful suggestion on terahertz transmission mechanism, and Mengxue Wang for cell culture and preparation.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *Ca-Cancer J. Clin.* **71**(1), 7–33 (2021).
2. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *Ca-Cancer J. Clin.* **71**(3), 209–249 (2021).
3. C. L. Chaffer and R. A. Weinberg, "A perspective on cancer cell metastasis," *Science* **331**(6024), 1559–1564 (2011).
4. A. W. Lambert, D. R. Pattabiraman, and R. A. Weinberg, "Emerging biological principles of metastasis," *Cell* **168**(4), 670–691 (2017).
5. C. Alix-Panabieres and K. Pantel, "Clinical applications of circulating tumor cells and circulating tumor DNA as liquid biopsy," *Cancer Discovery* **6**(5), 479–491 (2016).
6. J. Phallen, M. Sausen, and V. Adleff, *et al.*, "Direct detection of early-stage cancers using circulating tumor DNA," *Sci. Transl. Med.* **9**(403), eeaan2415 (2017).
7. E. Crowley, F. Di Nicolantonio, F. Loupakis, and A. Bardelli, "Liquid biopsy: monitoring cancer-genetics in the blood," *Nat. Rev. Clin. Oncol.* **10**(8), 472–484 (2013).
8. B. Hong and Y. Zu, "Detecting circulating tumor cells: current challenges and new trends," *Theranostics* **3**(6), 377–394 (2013).
9. M. Takao and K. Takeda, "Enumeration, characterization, and collection of intact circulating tumor cells by cross contamination-free flow cytometry," *Cytometry* **79A**(2), 107–117 (2011).
10. W. J. Allard, J. Matera, M. C. Miller, M. Repollet, M. C. Connelly, C. Rao, A. G. J. Tibbe, J. W. Uhr, and L. W. M. M. Terstappen, "Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases," *Clin. Cancer Res.* **10**(20), 6897–6904 (2004).
11. R. Rosenberg, R. Gertler, J. Friederichs, K. Fuehrer, M. Dahm, R. Phelps, S. Thorban, H. Nekarda, and J. R. Siewert, "Comparison of two density gradient centrifugation systems for the enrichment of disseminated tumor cells in blood," *Cytometry* **49**(4), 150–158 (2002).
12. S. Nagrath, L. V. Sequist, S. Maheswaran, D. W. Bell, D. Irimia, L. Ulkus, M. R. Smith, E. L. Kwak, S. Digumarthy, A. Muzikansky, P. Ryan, U. J. Balis, R. G. Tompkins, D. A. Haber, and M. Toner, "Isolation of rare circulating tumour cells in cancer patients by microchip technology," *Nature* **450**(7173), 1235–1239 (2007).
13. P. U. Jepsen, D. G. Cooke, and M. Koch, "Terahertz spectroscopy and imaging - modern techniques and applications," *Laser & Photon. Rev.* **5**(1), 124–166 (2011).
14. Y. Yoshida, X. Ding, K. Iwatsuki, K. Taniizumi, H. Inoue, J. Wang, K. Sakai, and T. Kiwa, "Detection of lung cancer cells in solutions using a terahertz chemical microscope," *Sensors* **21**(22), 7631 (2021).
15. H. Cheon, H. J. Yang, S. H. Lee, Y. A. Kim, and J. H. Son, "Terahertz molecular resonance of cancer DNA," *Sci. Rep.* **6**(1), 37103 (2016).

16. E. M. Hassan, A. Mohamed, M. C. DeRosa, W. G. Willmore, Y. Hanaoka, T. Kiwa, and T. Ozaki, "High-sensitivity detection of metastatic breast cancer cells via terahertz chemical microscopy using aptamers," *Sens. Actuators, B* **287**, 595–601 (2019).
17. Z. Zhang, G. Yang, F. Fan, C. Zhong, Y. Yuan, X. Zhang, and S. Chang, "Terahertz circular dichroism sensing of living cancer cells based on microstructure sensor," *Anal. Chim. Acta* **1180**, 338871 (2021).
18. K. Shiraga, Y. Ogawa, T. Suzuki, N. Kondo, A. Irisawa, and M. Imamura, "Characterization of dielectric responses of human cancer cells in the terahertz region," *J. Infrared, Millimeter, Terahertz Waves* **35**(5), 493–502 (2014).
19. M. Grognot and G. Gallot, "Quantitative measurement of permeabilization of living cells by terahertz attenuated total reflection," *Appl. Phys. Lett.* **107**(10), 103702 (2015).
20. Y. Zou, Q. Liu, X. Yang, H. C. Huang, J. Li, L. H. Du, Z. R. Li, J. H. Zhao, and L. G. Zhu, "Label-free monitoring of cell death induced by oxidative stress in living human cells using terahertz ATR spectroscopy," *Biomed. Opt. Express* **9**(1), 14–24 (2018).
21. Y. Cao, P. Huang, J. Chen, W. Ge, D. Hou, and G. Zhang, "Qualitative and quantitative detection of liver injury with terahertz time-domain spectroscopy," *Biomed. Opt. Express* **11**(2), 982–993 (2020).
22. Y. Sun, P. Du, X. Lu, P. Xie, Z. Qian, S. Fan, and Z. Zhu, "Quantitative characterization of bovine serum albumin thin-films using terahertz spectroscopy and machine learning methods," *Biomed. Opt. Express* **9**(7), 2917–2929 (2018).
23. H. Liu, K. Zhao, X. Liu, Z. Zhang, J. Qian, C. Zhang, and M. Liang, "Diagnosis of hepatocellular carcinoma based on a terahertz signal and VMD-CWSE," *Biomed. Opt. Express* **11**(9), 5045–5059 (2020).
24. O. Cherkasova, D. Vrazhnov, A. Knyazkova, M. Konnikova, E. Stupak, V. Glotov, V. Stupak, N. Nikolaev, A. Paulish, Y. Peng, Y. Kistenev, and A. Shkurinov, "Terahertz time-domain spectroscopy of glioma patient blood plasma: diagnosis and treatment," *Appl. Sci.* **13**(9), 5434 (2023).
25. X. Yang, M. Li, Q. Peng, J. Huang, L. Liu, P. Li, C. Shu, X. Hu, J. Fang, F. Ye, and W. Zhu, "Label-free detection of living cervical cells based on microfluidic device with terahertz spectroscopy," *J. Biophotonics* **15**(1), e202100241 (2022).
26. D. Hou, X. Li, J. Cai, Y. Ma, X. Kang, P. Huang, and G. Zhang, "Terahertz spectroscopic investigation of human gastric normal and tumor tissues," *Phys. Med. Biol.* **59**(18), 5423–5440 (2014).
27. L. Duvillaret, F. Garet, and J. L. Coutaz, "A reliable method for extraction of material parameters in terahertz time-Domain spectroscopy," *IEEE J. Select. Topics Quantum Electron.* **2**(3), 739–746 (1996).
28. D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science* **334**(6062), 1518–1524 (2011).
29. D. Cao, N. Xu, Y. Chen, H. Zhang, Y. Li, and Z. Yuan, "Construction of a pearson- and MIC-based co-expression network to identify potential cancer genes," *Interdiscip Sci Comput Life Sci* **14**(1), 245–257 (2022).
30. J. Das, J. Mohammed, and H. Yu, "Genome-scale analysis of interaction dynamics reveals organization of biological networks," *Bioinformatics* **28**(14), 1873–1878 (2012).
31. S. Yang, C. Li, Y. Mei, W. Liu, R. Liu, W. Chen, D. Han, and K. Xu, "Discrimination of corn variety using terahertz spectroscopy combined with chemometrics methods," *Spectrochim. Acta, Part A* **252**, 119475 (2021).
32. K. Li, X. Chen, R. Zhang, and E. Pickwell-MacPherson, "Classification for glucose and lactose terahertz spectrums based on SVM and DNN methods," *IEEE Trans. Terahertz Sci. Technol.* **10**(6), 617–623 (2020).
33. M. S. Tehrany, B. Pradhan, and M. N. Jebur, "Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS," *J. Hydrol.* **512**, 332–343 (2014).
34. B. Born and M. Havenith, "Terahertz Dance of Proteins and Sugars with Water," *J. Infrared, Millimeter, Terahertz Waves* **30**(12), 1245–1254 (2009).